



Safer and responsible AI systems via Policy and Platforms

Ramayya Krishnan

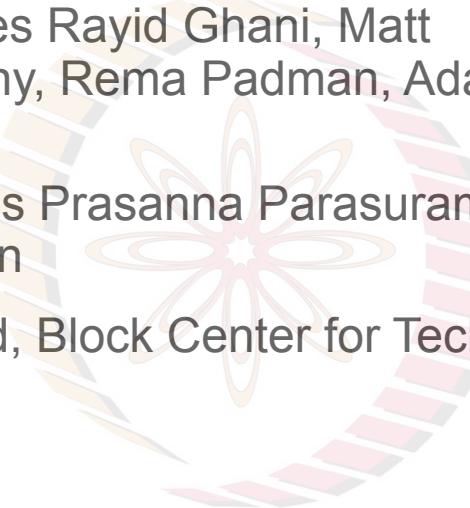
Dean, **Heinz College Of Information Systems And
Public Policy**

Founding Faculty Director, **Block Center for
Technology and Society**

IITM Safe and RAI Summer School, June 2024

Acknowledgements

- Thanks to my CMU colleagues Rayid Ghani, Matt Fredrickson, Lauren McIlvenny, Rema Padman, Adam Papper
- Thanks to my NYU colleagues Prasanna Parasurama, Joao Sedoc and Arun Sundararajan
- Data Science for Social Good, Block Center for Technology and Society, Metro 21



NPTEL

Block Center for Technology & Society



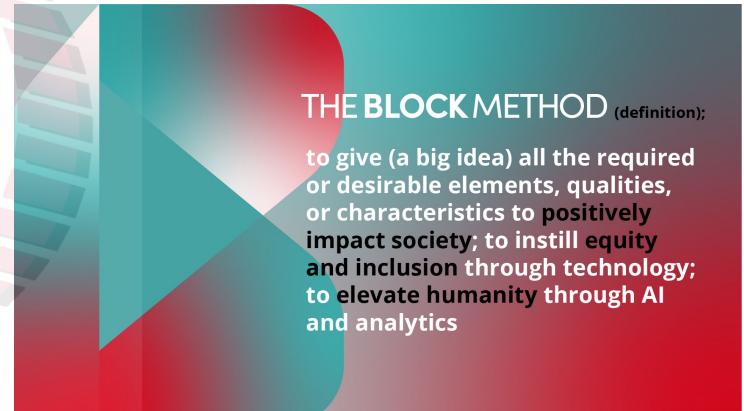
The Future of
Work



Responsible AI



NPTEL



THE BLOCK METHOD (definition):

to give (a big idea) all the required or desirable elements, qualities, or characteristics to positively impact society; to instill equity and inclusion through technology; to elevate humanity through AI and analytics

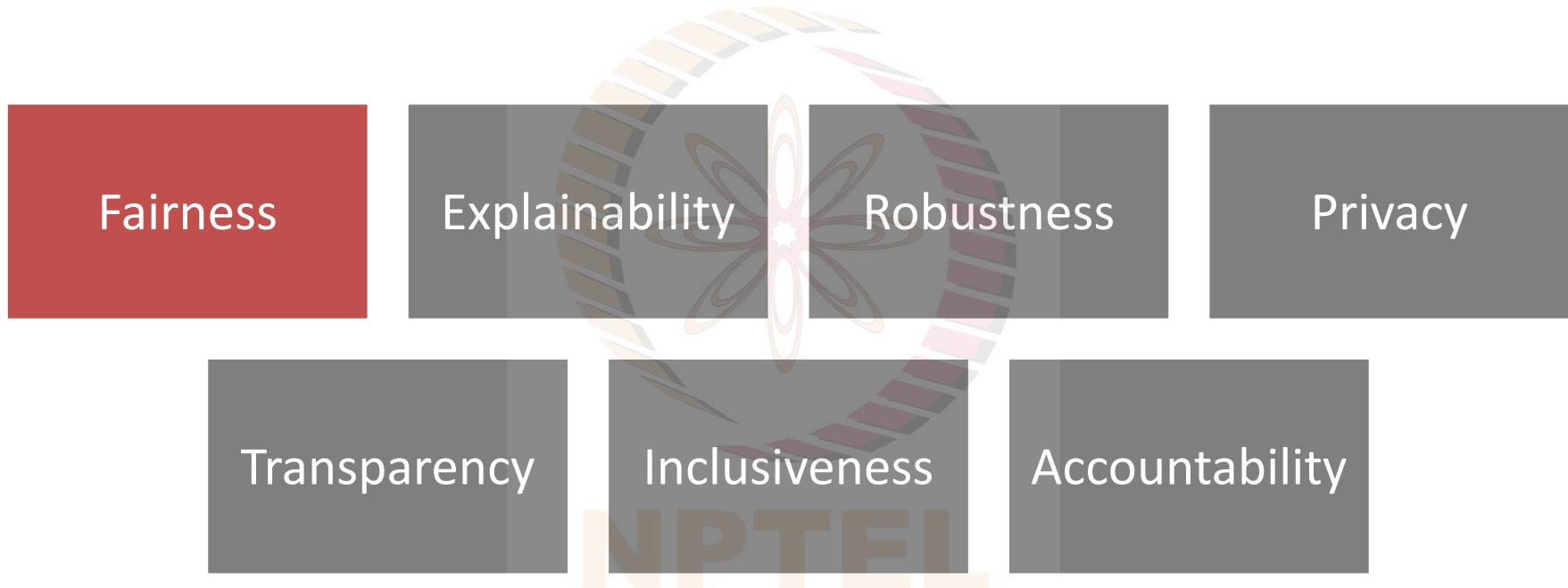


Carnegie Mellon University
Responsible AI

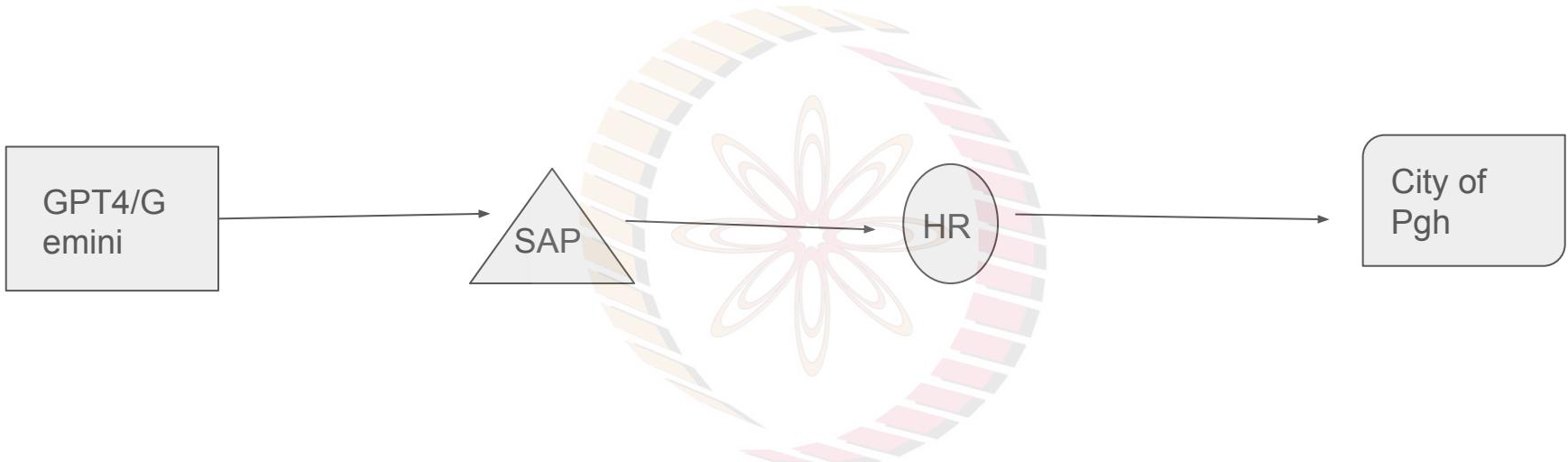


The Block C
FOR TECHNOLOGY AI

RAI Principles



An HR Application - the AI value chain



Data and information flow up and down the value chain. Evaluation of the “system” vs. just the AI. What about obligations? of the Model Developer, the Deployer and the User?

NPTEL

From Policy to Practice

- Measurement and tools to evaluate an AI model and AI system on each of these dimensions of responsible AI
- Role of these in procurement (most organizations will procure AI for consequential applications)
 - Contrast with office 365 or Google Docs which will integrate AI co-pilots for productivity
- See new NTIA report on accountable AI
(https://www.ntia.gov/sites/default/files/publications/ntia_ai_report_final-3-27-24.pdf)

Key Areas of Policy Interest

AI and Democracy: Deep Fakes, Synthetic media

- **Code of conduct, standards, detection technology**
- **Synthid, c2pa.org**

AI Safety and security (pre-deployment)

- **Model transparency and auditing**

AI Safety and Security (post deployment)

- **CERT for AI**

NPTEL

Two Executive Orders on AI



September 20, 2023



October 30, 2023

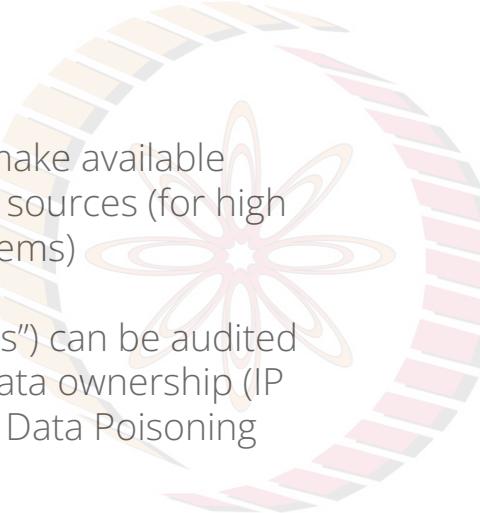
Pre-Deployment AI safety and security



NPTEL

Model Transparency: Input Data Cards and Analysis

- Builders of AI models must make available information about their data sources (for high risk and safety critical AI systems)
- These inputs (“nutrition labels”) can be audited to ensure compliance with data ownership (IP and copyright), Bias analysis, Data Poisoning attacks etc.
- Example: GPT 4 System Card
<https://cdn.openai.com/papers/gpt-4-system-card.pdf>



Nutrition Facts
8 servings per container
Serving size 2/3 cup (55g)

Amount per 2/3 cup
Calories **230**

% DV*

12%	Total Fat 8g
5%	Saturated Fat 1g
	Trans Fat 0g
0%	Cholesterol 0mg
7%	Sodium 160mg
12%	Total Carbs 37g
14%	Dietary Fiber 4g
	Sugars 1g
	Added Sugars 0g
	Protein 3g
10%	Vitamin D 2mcg
20%	Calcium 260mg
45%	Iron 8mg
5%	Potassium 235mg

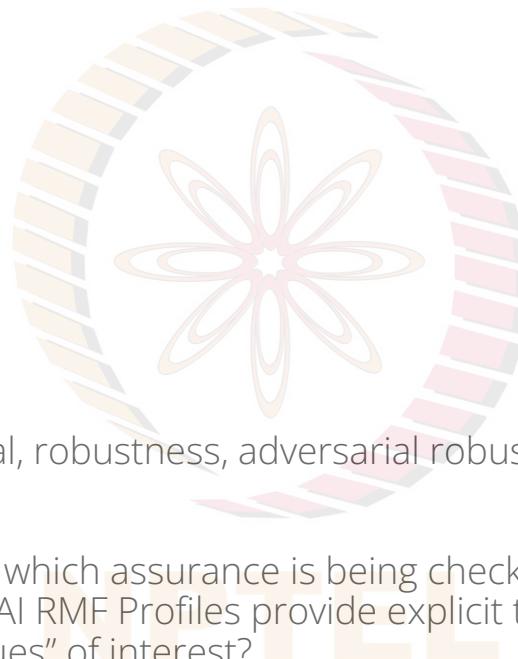
* Footnote on Daily Values (DV) and calories reference to be inserted here.

Model Transparency: The AI Underwriters lab

- Our tolerance for errors will vary by use. Systems that protect life and limb should allow for very few mistakes regardless of the cost. Eg. air traffic control vs. an email spam filter
- We propose that mechanisms should be created to develop risk standards (expressed as AI RMF profiles) and auditors can assess if the AI system complies with the risk thresholds for safety critical and high risk applications
- Like FASB for accounting standards and Underwriters Lab



Challenges to be addressed



Documentation and processes

- Ethics by design

Measurement and metrics

- Operationalizing values
 - Codifying tradeoffs
 - Thresholds for capabilities

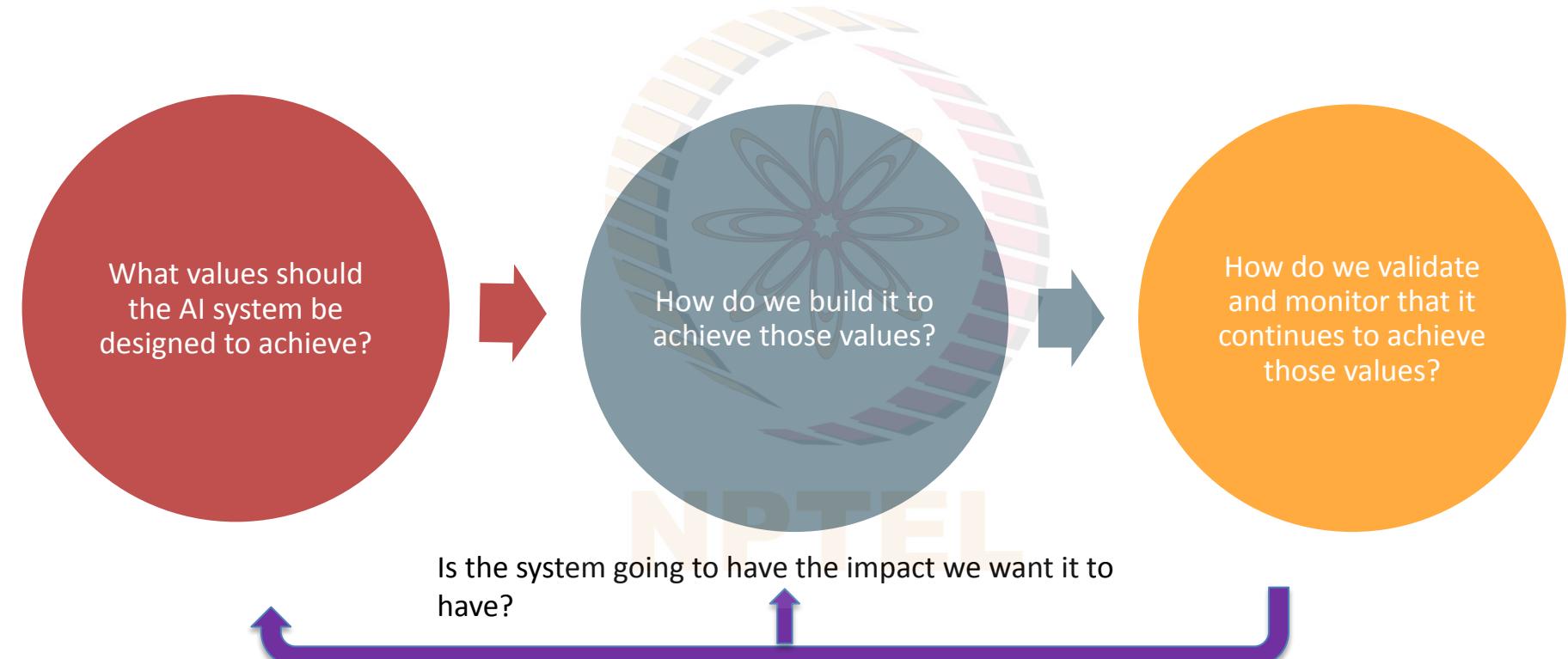
Evaluation

- Capabilities (knowledge, retrieval, robustness, adversarial robustness, reliability, accuracy, equity,..)

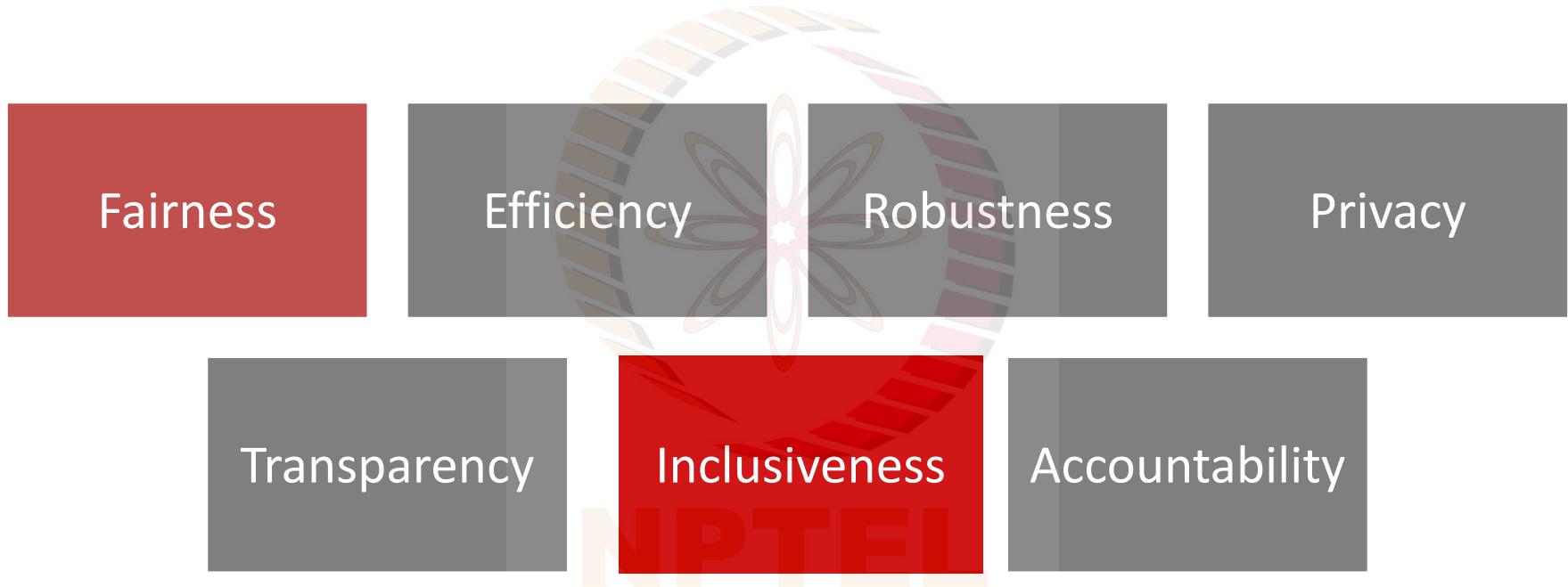
Assurance

- What are the standards against which assurance is being checked?
 - What is the GAAP for AI? Could AI RMF Profiles provide explicit thresholds for what is deemed acceptable for the multiple “values” of interest?

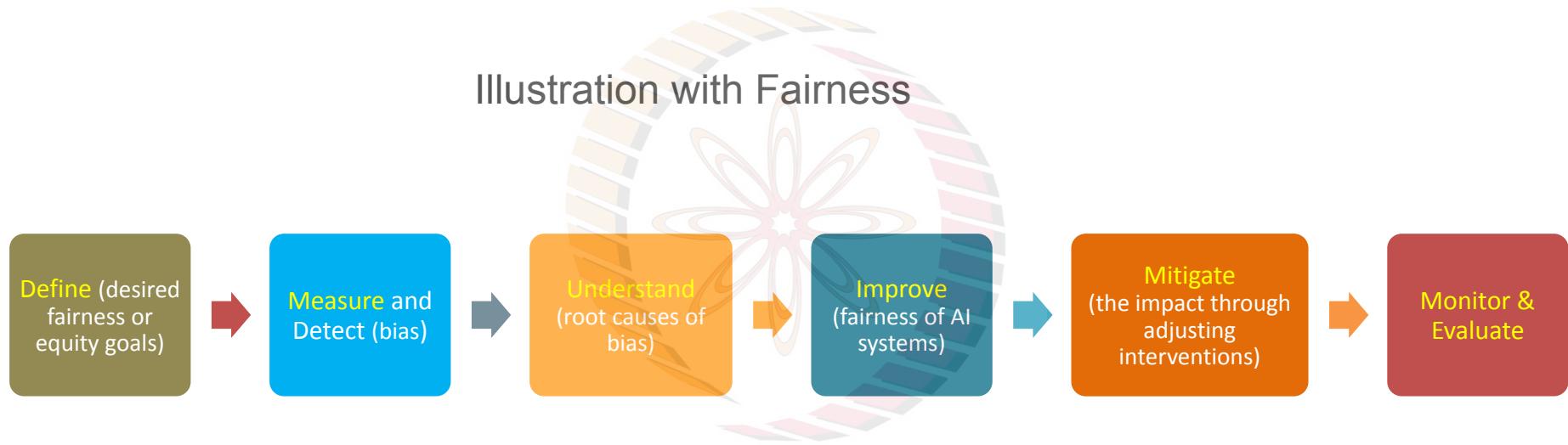
How should we design, develop, and deploy AI systems?



What values should we design for?



Adopting a Systems Approach to AI design and development - focus on the system rather than just the AI model



NPTEL

Societal USE Cases

Increasing Educational Outcomes in Schools (10+ school districts across the US and with Department of Education, El Salvador)



Matching interventions to students in need of extra support

A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. Lakkaraju et al. KDD 2015



Reducing Health and Safety Issues in Rental Housing

Reducing number of people going to Jail (Johnson County, KS)
Reducing Incarceration through Prioritized Interventions. Bauman et. Al. ACM COMPASS 2018

11 MILLION
people move through 3,100 Jails

\$22 BILLION
in cost

64%
suffer from mental illness

68%
have a substance abuse disorder

44%
suffer from chronic health problems

Cycle of Incarceration



DONORS
CHOOSE

Find a classroom to support

About us

Help



**Support a classroom.
Build a future.**

Teachers and students all over the U.S. need your help to bring their classroom dreams to life. Get crayons, books, telescopes, field trips, and more for a classroom today.

[See classroom projects](#)

Our efficiency and transparency have earned us the highest rating on Charity Navigator.

<http://www.donorschoose.com/>

Predict priority and allocate scarce resource pattern

Table 1: Policy Settings and Data Details

	Inmate Mental Health	Housing Safety	Student Outcomes	Education Crowdfunding
Prediction Task	Jail booking within the next 12 months	Housing unit having a violation within the next year	Student not returning to school next year	Project not getting fully funded within 4 months
Timespan	2013-01-01 to 2019-04-01	2011-01-01 to 2017-06-01	2009-01-01 to 2018-01-01	2010-01-01 to 2014-01-01
# of Entities	61,192	4,593	801,242	210,310
# of Features	3,465	1,657	220	319
Base Rate	0.12	0.43	0.25	0.24
Evaluation Metric	Precision at top 500	Precision at top 500	Precision at top 10,000	Precision at top 1,000
Sensitive Attribute	Race	Median Income	Age Relative to Grade	Poverty Level

Measuring Fairness

- Variety of measures
- Makhlouf et al. (2020) (2022), Rodolfa et al. (2020)
- Group measures, individual measures
- Fairness Tree (Ghani et al.)
- Also what about process fairness in addition to output fairness?



NPTEL

Fairness Tree Framework

Are your interventions
punitive or assistive?

Punitive
(could hurt individuals)

Assistive
(will help individuals)

Can you intervene with
most people with need
or only a small fraction?

Small Fraction

Most People

Among which group are you
most concerned with ensuring
predictive equity?

Everyone w/o regard
for actual outcome

People for whom
intervention is taken

Intervention
NOT warranted

FP/GS Parity

FDR Parity

FPR Parity

Recall Parity*

Everyone w/o regard
for actual need

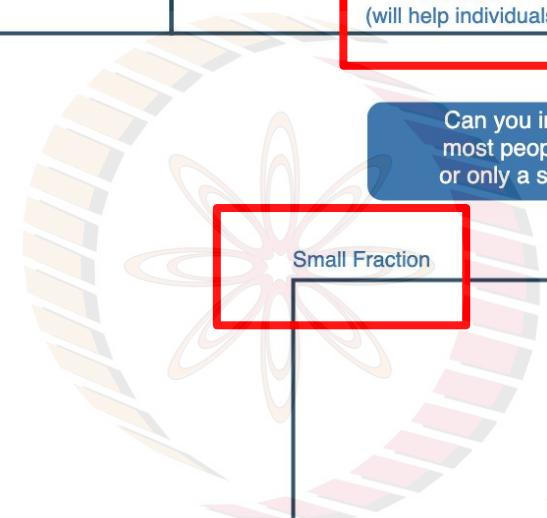
People NOT
receiving assistance

People with
actual need

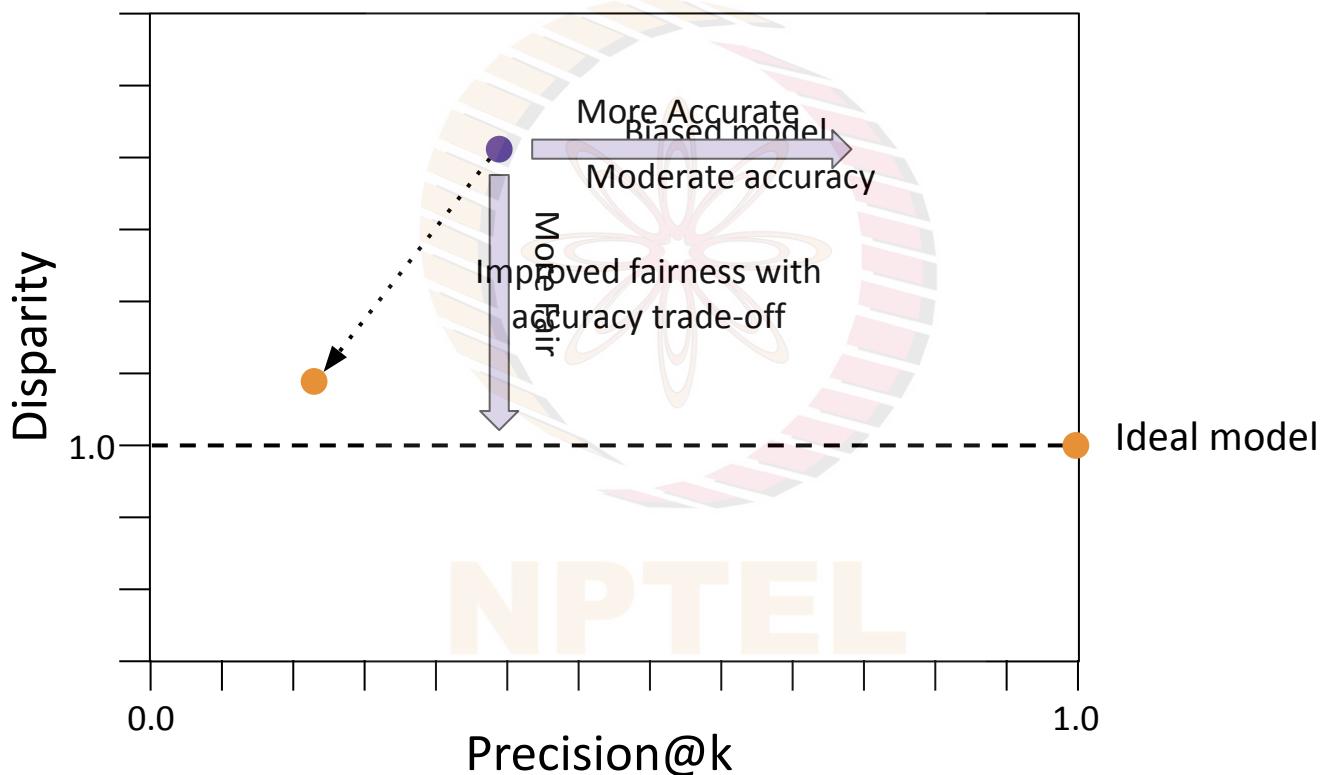
FN/GS Parity

FOR Parity

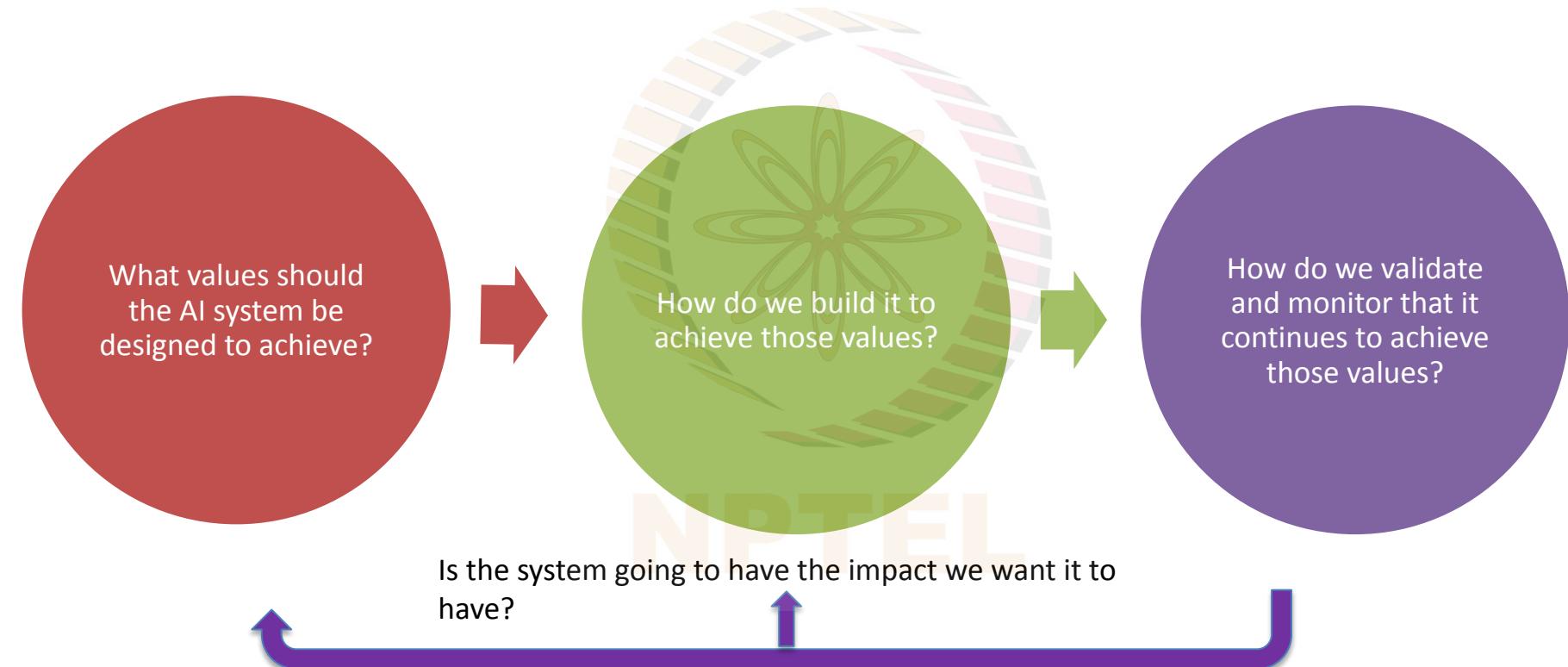
FNR Parity



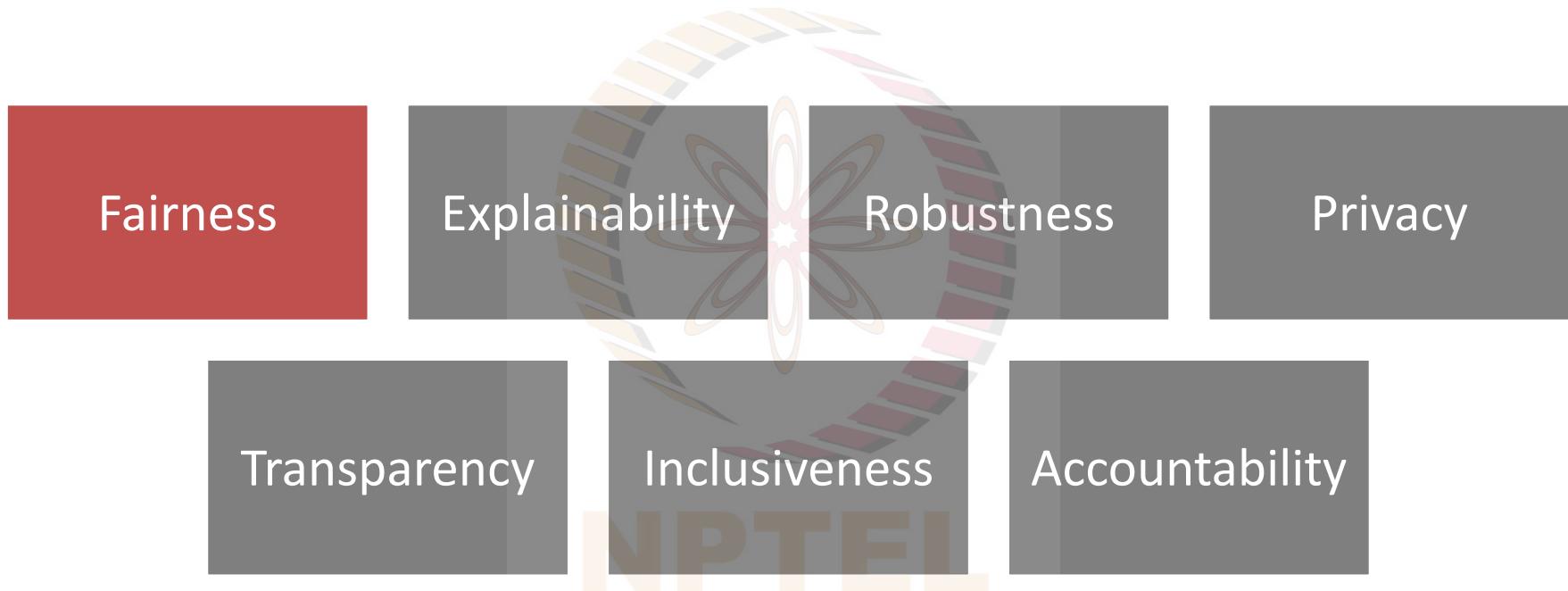
Accuracy vs. Disparity



How should we design AI systems?



What values should we design for?



Health Literacy: Retrieving quality understandable content



How to Use an Insulin Pen - Mayo Clinic Patient Education



Mayo Clinic 1.02M subscribers

Subscribe

4.9K



Share

Download



Framework to Assess Healthcare Video Understandability

A video is understandable when

- Consumers of diverse backgrounds and varying levels of health literacy can process and explain key messages

<https://www.ahrq.gov/health-literacy/patient-education/pemat1.html>

- Evaluation of patient educational videos have relied on the judgment of domain experts on several critical dimensions (Backinger et al. 2011)
 - Content understandability by end users (Ruppert et al. 2017)
 - The volume of medical information (Liu et al. 2020)
 - The complexity of medical information provided (Stellefson et al. 2014)
- Agency for Healthcare Research and Quality (AHRQ) proposed the Patient Education Materials Assessment Tool (PEMAT) (Shoemaker et al. 2014)
 - Evaluates and compares patient education materials in written, audio and video formats
 - PEMAT highlights the need to emphasize the understandability of patient educational materials (Liu et al. 2020, 2023)

Patient Educational Material Assessment Tool – Video Understandability

Content Scoring: 0 = disagree, 1 = Agree, N/A = Not applicable

1	The material makes its purpose completely evident.	0, 1
---	--	------

Word Choice & Style

2	The material uses common, everyday language.	0, 1
3	Medical terms are used only to familiarize audience with the terms. When used, medical terms are defined.	0, 1
4	The material uses the active voice.	0, 1

Organization

5	The material breaks or “chunks” information into short sections.	0, 1, N/A
6	The material’s sections have informative headers.	0, 1, N/A
7	The material presents information in a logical sequence.	0, 1
8	The material provides a summary.	0, 1, N/A

Layout & Design

9	Text on the screen is easy to read.	0, 1, N/A
10	The material allows the user to hear the words clearly (e.g., not too fast, not garbled).	0, 1, N/A
11	The material uses illustrations and photographs that are clear and uncluttered.	0, 1, N/A
12	The material uses simple tables with short and clear row and column headings.	0, 1, N/A

$$\left(\text{Understandability} = \frac{\text{The total number of 1's in PEMAT result}}{12 - \text{the total number of NA's in the PEMAT result}} \times 100\% \right)$$



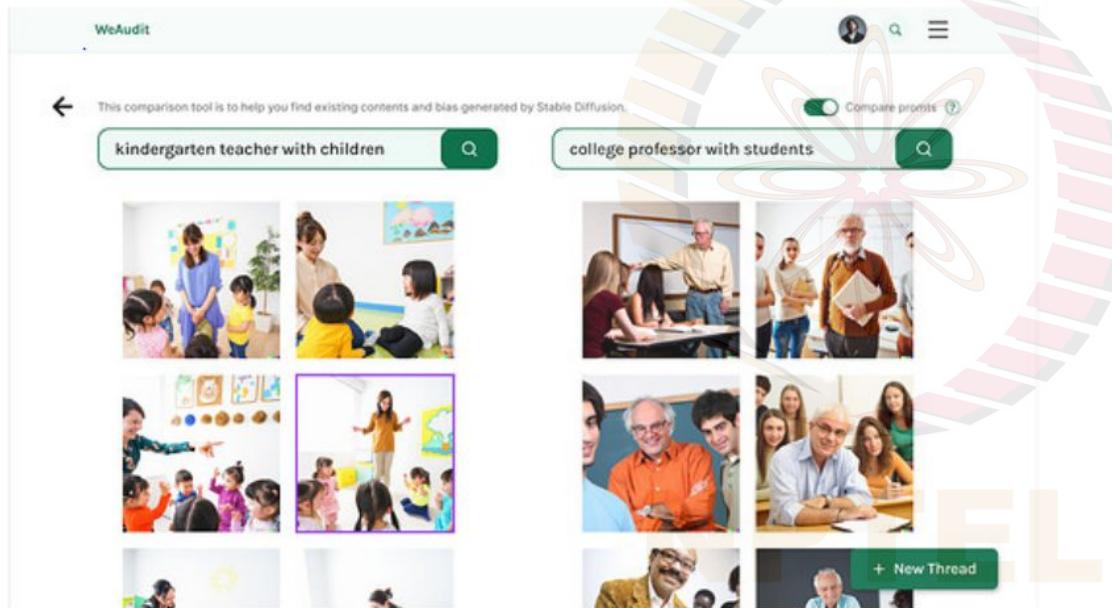
Features for Video Understandability Classification

(Liu et al. 2020, 2023)

Classifier F1 based on Video Metadata		al. 2020, 2023	Classifier F2 based on Video Content Analysis
The material makes its purpose evident	<ul style="list-style-type: none">• Has title• Has video description• Has tags	Use common everyday language	Automated readability index (Webster 2019; McCallum et al. 1982)
Use common everyday language	Automated readability index	Materials uses active voice	The number of verbs in active voice in the video transcripts
Material uses active voice	The number of verbs in active voice in the video description	Video sections have informative title	Text recognition in videos
The material provides a summary	The number of summary words in video description	The material provides a summary	The number of summary words in video transcripts
Other features	<ul style="list-style-type: none">• Video duration• The total # of words in the video description• The total # of unique words in the video description• Total # of sentences in the video description• The total # of medical terms in the video description	Materials may allow users to hear the words clearly Text on screen is easy to read • Other features	video transcription confidence scores text recognition confidence score <ul style="list-style-type: none">• Total # of words in the video transcript• Total # of unique words in video transcript• Total # of sentences in the video transcript



weaudit.org (Hong and Parer, CMU)



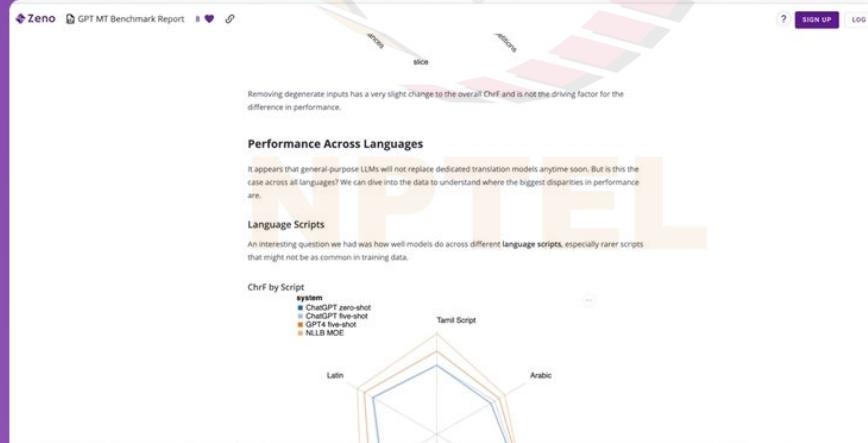
Example: kindergarten teacher VS college professor

Here we can see that all images generated for kindergarten teacher are female whereas all generated for a college professor are male.

AI Evaluation Made Easy

Discover how your AI performs with **Zeno**.

Explore your data, uncover failures, and create beautiful, interactive charts.

[Get Started](#)[Explore Projects](#)

Harms and impact

Who bears of the cost of these errors?

Is it borne disproportionately by disadvantaged groups?

How can these be assessed prior to deployment?

The NPTEL logo is a watermark located at the bottom center of the slide. It consists of the word "NPTEL" in a large, bold, sans-serif font, with each letter in a different color: N is orange, P is light blue, T is green, E is red, and L is yellow. Below the letters is a stylized orange graphic element resembling a flame or a series of overlapping curved lines.

Post Deployment AI Safety and Security



NPTEL

AI Lead Emergency Response Team (ALRT)

- Even with due diligence pre-deployment, need to respond to failures and vulnerabilities discovered post deployment
- Catalog failures, vulnerabilities and incidents, forensic analysis and risk tiering and sharing of data upstream to developers and downstream to deployers
- Similar to CMU's Cyber Emergency Response Team (CERT). Needed for .mil, .com and .gov domains



Suffix attacks (work of Fredrikson et al.)

Chatbots are “aligned” to a policy when they are trained



They are good at following these policies under normal circumstances
...but fail when probed *adversarially*

Attacking (proprietary) public LLMs



Step 1:

Find a *universal* attack that breaks several open-source LLMs



Step 2:

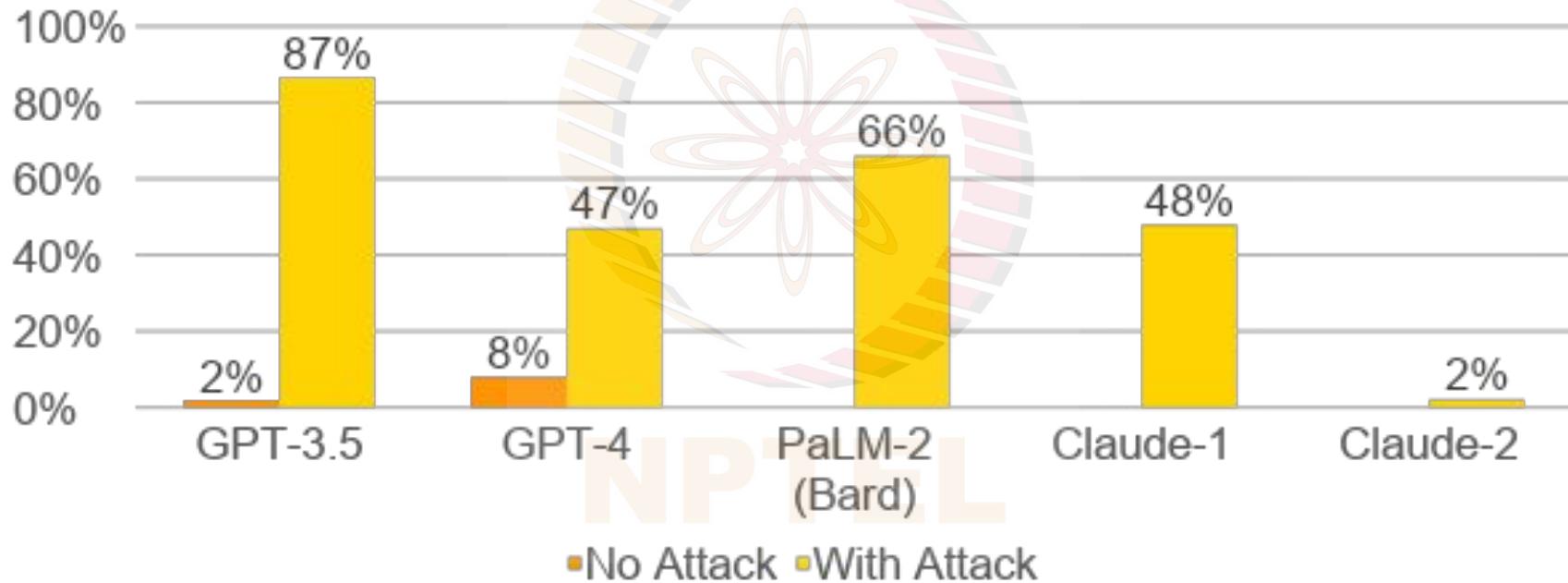
Paste the attack into public chatbots



Step 3:

... there is no Step 3...

Attack Success Rate on Black-Box Models



Is this significant?



Language models are being used fundamentally differently than previous forms of machine learning

1. Replacement for human-coded functionality
2. Autonomous decision-making
3. ...

This vulnerability lets an attacker wrest control of an AI component, disregarding common guardrails

Closing thoughts

Much to learn about how to evaluate frontier AI models

Research needed to develop a science of measurement, evaluation and assurance of AI systems

Multi-disciplinary approaches are required that draw on social science, engineering, and computer science to respond to this challenge